Visualising the Propagation of News on the Web

Svitlana Vakulenko*, Max Göbel†, Arno Scharl* and Lyndon Nixon*

* MODUL University Vienna
† Vienna University of Economics and Business
Vienna, Austria
{svitlana.vakulenko,arno.scharl,lyndon.nixon}@modul.ac.at
max.goebel@wu.ac.at

Abstract

When newsworthy events occur, information quickly spreads across the Web, along official news outlets as well as across social media platforms. Information diffusion models can help to uncover the path of an emerging news story across these channels, and thereby shed light on how these channels interact. The presented work enables journalists and other stakeholders to trace back the distribution process of news stories, and to identify their origin as well as central information hubs who have amplified their dissemination.

1 Introduction

Newsworthy events are communicated via traditional news media sources such as CNN and the New York Times, as well as social media platforms. However, the specific path a story takes via various news distributors and the interplay with the social network discussion is not well studied yet. This limits further research on rumour detection and news content verification. This paper presents an approach developed in the EU-funded PHEME project (www.pheme.eu), tracking information contagions across various media sources including major online news publishers as well as single Twitter users.

In: M. Martinez, U. Kruschwitz, G. Kazai, D. Corney, F., R.and D. Albakour (eds.): Proceedings of the NewsIR'16 Workshop at ECIR, Padua, Italy, 20-March-2016, published at http://ceur-ws.org

2 Related Work

Information diffusion is an established research field traditionally applied to explicit networks such as social media, but less studied in communication scenarios where information sources tend to be implicit.

One research area that links news articles to trace the origin of an information piece is text reuse (plagiarism) detection. This approach has been recently applied to analyse information exchange networks based on historical newspaper texts [CIK14] and to study the evolution of memes [SHE+13]. In contrast to this work, our approach does not track stable phrases, but uses information pieces directly as relations.

Yang and Leskovec [YL10] model the total number of infected nodes over time determined by the influence function of nodes infected in the past. They formulate this problem as an instance of *Non-Negative Least Squares* and use it to predict the volume of information diffusion in the future. Their approach differs from ours since it does not model implicit network to surface implicit links between the information sources.

3 Information Diffusion Model

3.1 Modeling Information Contagions

We propose a 'bag-of-relations' document representation model to capture the essential information contained in textual documents, such as news articles. The main idea behind our approach is to represent each document as a set of relations, represented as n-grams-like similarity strings. Unlike n-grams, these strings are constructed from grammatical dependency relations instead of the sequential order of words in a sentence. We employ a dependency parser to obtain parse trees for each of the sentences and extract the relations by traversing these trees. The relations are then modeled as triples of the form:

s (subject) – **p** (predicate) – **o** (object)

We start off with the task of finding all the predicates in the sentence, which play the role of triggers to finding the corresponding relations. We normalize the predicates to the form: '{*synsets (or lemmas)*} + {*flags*}', by detecting for each verb the corresponding WordNet synset (or taking the verb's lemma otherwise), tense, voice, negation and auxiliary verbs (e.g. 'did not say' is transformed to 'state D N').

We define a set of words to be excluded from the predicate phrase to improve the results. For example, there are trivial relations, which are common among all news articles and which we would like to eliminate, e.g. the ones triggered by the predicates: 'print', 'post', 'update'. Words that do not carry any semantic information of the predicate, but are used solely for grammatical purposes (e.g. 'will', 'do'), are also excluded.

We introduce special symbols to preserve the grammatical information removed at the previous step. As such, D indicates the past tense, F – future tense, N – negation, A - auxiliary verb ('would'). Since there are multiple ways to express negation or past tense, this approach allows to disambiguate and group together semantically-equivalent relations.

Then, for each predicate we pick the adjacent branches with clauses that correspond to the subject and objects of the relation. We designed a simple heuristic for English language texts: assign the node to the subject-element if it precedes the predicate in the sentence, and to the object otherwise (i.e. when it follows the predicate).

We construct separate relations for each object-element related to the predicate and one relation with an empty object, if the subject is not empty. This simple heuristic allows us to create several fine-grained relations with different levels of detail. For example, a sentence "The plane landed in Panama on Tuesday" will be decomposed into: 'plane - land D', 'plane - land D - in Panama', 'plane - land D - on Tuesday'. This approach enables us to spot those articles that report on the same event but provide complementing or contradicting details.

3.2 Modeling Diffusion Cascades

We assume that all articles sharing the same information contagion are related to each other, i.e there is a path for every pair of articles within the diffusion graph. We included this assumption into our model by enforcing the connectivity requirement over our diffusion graph: for each node (except the root node), we generate an incoming edge that links the node to its source. Here, we also use the single source assumption: for all nodes (except the root node), there is exactly one incoming edge linking the node to its source (the closest neighbour). This assumption allows us to simplify the model and avoid making assumptions about the similarity threshold value, i.e. how similar the articles should be to be linked in the diffusion model.

The diffusion process is modeled as a graph with two types of edges: (1) explicit links referencing the source URL - edge direction: from the source to the post with the URL; (2) implicit links to connect similar posts that share the same information contagions - edge direction: from the older to the more recent post.

We link news articles to social media posts by querying the Twitter API with the URL of a news article to obtain all the tweets which reference it explicitly. News media often do not cite their information sources apart from the references to the major news agencies, e.g. Reuters. Therefore, we focus on uncovering the latent relations between the news articles, which we construct based on content similarity. We construct the diffusion graph with edges generated using the pairwise similarity values computed over the relation bags of the articles.

There are two methods to compute similarity between a pair of news articles: (1) considering the intersection of the relation bags, (2) hashing the relation bags and computing the similarity between the relation hashes. While the first method, returning an integer for the number of shared relations, is simple and intuitive, it is limited to considering only exact matches between relations. The second method is more powerful by allowing for approximately similar relations.

We test both methods to compute similarity for any two relation bags complementary to each other to evaluate which of them performs better in practice. We use Nilsimsa hashing and Hamming distance to generate and compare the relation hashes. Nilsimsa is one of the most popular algorithms for locality-sensitive hashing and is traditionally employed for spam detection in emails. Hamming distance measures the proportion of positions in the hash at which the corresponding symbols are different.

4 Experiment

4.1 Dataset and Configuration

The dataset is based on a recent news media snapshot exported from the PHEME dashboard [SWG+16], which contains 71,000 articles published between 27 November and 3 December 2015. We ran the relation extraction procedure on this corpus and picked one of the frequent information contagions to illustrate how it can be backtracked across the online media:

s: president barack obama – **p**: state D – **o**:

This relation provided us with a cluster of 12 news articles. It is able to capture all the expressions with

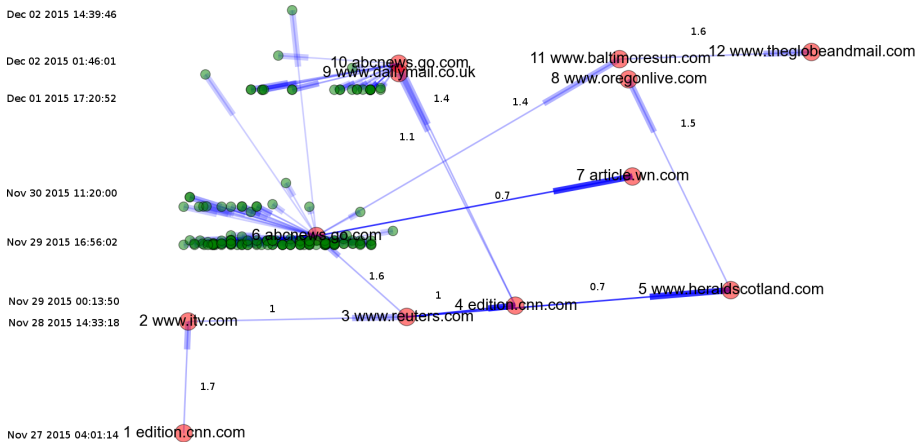

Figure 1: Sample information diffusion model: 'president barack obama state D'

a predicate that belongs to the WordNet synset 'state' and is used in the past tense ('D'), such as "president barack obama said", thereby indicating statements made by President Obama.

For each article we retrieved the tweets via its URL using Twitter Search API, which resulted in 150 tweets (127 and 23 for two of the articles). We used the networkx[1] and matplotlib[2] Python libraries to visualize the resulting diffusion graph (see Figure 1).

4.2 Results

The nodes of the graph in Figure 1 represent the individual posts published at discrete time intervals (red: news articles; green: tweets). The sources get infected in the sequence order aligned along the vertical time axis as indicated on the node labels. The same source may appear more than once within the same network if it has published multiple articles containing the same information contagion within the given time interval.

The edges of the graph represent *direct links* in case of tweets, or *content similarity* in case of articles. Content similarity values are indicated as weights over the corresponding edges. Values closer to 0 indicate more similar articles. *Light edges* indicate that the adjacent articles share a single information contagion, while *solid edges* indicate that the articles have more than one information contagion in common.

5 Conclusion and Future Work

We showed how to uncover the latent relations between news articles and used them to infer a model of the implicit diffusion network, which constitute an important step towards rumour detection research. The results of our initial experiment indicate that our relation-based modeling approach is promising and merits further research. In future work we will further evaluate our approach against baseline methods.

6 Acknowledgments

The presented work was conducted within the PHEME Project (www.pheme.eu), which has received funding from the European Union's Seventh Framework Programme for Research, Technological Development and Demonstration under Grant Agreement No. 611233.

References

[CIK14] Giovanni Colavizza, Mario Infelise, and Frederic Kaplan. Mapping the Early Modern News Flow: An Enquiry by Robust Text Reuse Detection. In *Social Informatics*, pages 244–253, 2014.

[SHE+13] Caroline Suen, Sandy Huang, Chantat Eksombatchai, Rok Sosic, and Jure Leskovec. NIFTY: A System for Large Scale Information Flow Tracking and Clustering. In *22nd International Conference on World Wide Web*, pages 1237–1248, 2013.

[SWG+16] Arno Scharl, Albert Weichselbraun, Max Göbel, Walter Rafelsberger, and Ruslan Kamolov. Scalable knowledge extraction and visualization for web intelligence. In *49th Hawaii International Conference on System Sciences*, pages 3749–3757, 2016.

[YL10] Jaewon Yang and Jure Leskovec. Modeling information diffusion in implicit networks. In *10th International Conference on Data Mining*, pages 599–608, 2010.

[1] networkx.github.io
[2] matplotlib.org